MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 1, 2018/2019

TPA7021 – DATA PRE-PROCESSING AND ANALYSIS

(All Sections / Groups)

3 October 2018
8:00 p.m. – 10:00 p.m.
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This paper consists of 5 pages (including cover page) with 5 Questions only.

2. Attempt **ALL FIVE** questions. The distribution of the marks for each question is given. This paper carries **50 marks**. Marks will be pro-rated to reflect 40% of the final examination for the course.

3. Please print all your answers in the Answer Booklet provided.

Question 1 (10 Marks)

a) What is the definition of Knowledge Discovery in Databases (KDD)? You can use your own words to describe it.

[1 mark]

b) The following are the 6 common stages of a Knowledge Discovery and Data Mining (KDD) process.

- Problem Specification.
- Problem Understanding.
- Data Preprocessing.
- Data Mining.
- Evaluation.
- Results Exploitation.

Give a brief explanation for each of the stages.

[6 marks]

c) Data reduction can be achieved through "Feature Selection", "Instance Selection" and "Discretization" methods. Explain with the assistance of diagrams how these methods can obtain a reduced representation of the original data .

[3 marks]

Question 2 (10 Marks)

a) State the SIX (6) forms of data preparation methods and provide a brief explanation for each.

[6 marks]

b) In partitioning a dataset, two main problems may arise by using the same data to train and evaluate the model. These are over-fitting and under-fitting. Explain what is meant by over-fitting and under-fitting.

[1 mark]

Continued ...

c) Explain the data partitioning process "*5x2 CV*". (You may use diagrams to assist your explanation).

[3 marks]

Question 3 (10 Marks)

a) The following conditions are needed in order to safely carry out parametric tests: Independence, Normality and Heteroscedasticity. Give a brief explanation of the 3 conditions mentioned.

[3 marks]

b) The FOUR (4) basic steps of data preparation are:

- Data Integration
- Data Cleaning
- Data Normalization
- Data Transformation

Explain each of these steps.

[4 marks]

c) In data normalization, if the minimum or maximum values of an attribute is not known, or the data is noisy, the min-max normalization method is infeasible. Why is it not feasible to use min-max normalization?

[1 mark]

d) In the absence of the minimum or maximum values of an attribute, we can use Z-score normalization. Explain how the Z-score normalization works.

[2 marks]

Continued ...

Question 4 (10 Marks)

a) What are the three missing value mechanism assumptions? Give a brief explanation on the assumptions that lead to the missing values.

[3 marks]

b) What is the purpose of conducting the data reduction process?

[2 marks]

c) Data sampling is one of the methods used to reduce the number of instances submitted to the Data Mining algorithm. Give a brief description of the following FIVE (5) data sampling methods:

- Simple random sample without replacement
- Simple random sample with replacement
- Balanced sample
- Cluster sample
- Stratified sample

[5 marks]

Question 5 (10 Marks)

a) Feature Selection is a process that chooses an optimal subset of features according to a certain criterion. State and explain at least THREE (3) purposes of conducting the Feature Selection process.

[3 marks]

b) Differentiate between Instance Selection and Data Sampling.

[2 marks]

c) In what situation would Instance Selection be preferred than Data Sampling?

[1 mark]

Continued ...

d) The discretization process transforms quantitative data into qualitative data, that is, numerical attributes into discrete or nominal attributes with a finite number of intervals, obtaining a non-overlapping partition of a continuous domain. What are the FOUR (4) advantages of using discretization?

[4 marks]

END OF EXAM PAPER